UNIVERSIDAD
COMPLUTENSE
MADRID

FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

GRADO EN ECONOMÍA
TRABAJO DE FIN DE GRADO

TÍTULO: Evolution of the wage gap in Spain by age and gender between 2006 and 2014.

AUTOR: Blanca Rivera Garrido

TUTOR: Jesús Barreal Pernas

CURSO ACADÉMICO: 2018/2019

CONVOCATORIA: Junio

Index

ABSTRACT

The gender wage gap is a relevant policy topic affecting our societies. This paper analyzes the gender wage gap in Spain depending on the age —and its evolution— by using quadrennial microdata provided by the *Instituto Nacional de Estadística* for three different years, 2006, 2010, and 2014. Our results show that the gap still exists and is likely to have increased due to the crisis. Female workers struggle with getting job positions with similar wages compared to men and the situation does not seem to be improving. There is a particularly unfair gender wage gap: a woman getting a lower salary compared to a man when working exactly in the same job —what we call arbitrary gender wage gap. Our findings in the three years analyzed suggest that this arbitrary gender wage gap, although unbearable, tends to decrease for younger cohort of workers. In general women have struggled to reconcile their time between work and family life; the measures taken to reduce this gap are not likely to benefit as much older women who have already been affected. However, further research is needed to assess the causes of this trend and to better understand why older female workers are left behind.

ACKNOWLEDGMENTS

I would like to thank the R Development Core Team (R CoreTeam, 2018) for providing the tool that enables this quantitative analysis, the developers of the R-studio environment (RStudio Team, 2016), and all the developers of the libraries used in the analysis (Hlavac, 2018b; Wickham, 2016; Zeileis & Hothorn, 2002; Zeileis & Grothendieck, 2005).

CHAPTER 1. INTRODUCTION

The wage gap is a well-known issue with relevant socioeconomic implications. There is consensus that there are some inequalities in the labor force between women and men, but when looking at the details of these inequalities, discrepancies arise. To compound the problem, people tend to capture this gap through a figure which can be misleading (Kessler, 2014; Rosin, 2013).

This paper is going to analyze how the wage gap between females and males varies among age groups. In the US, occupational segregation among age groups showed an overall decrease among all ages with the youngest age group leading the way. This transition went hand in hand with changes in educational achievements. But in the 1990s this progressed stalled and started increasing for younger women (Hegewisch, Liepmann, Hayes, & Hartmann, 2010). In Europe, the wage gap is lower for younger workers and it increases for older workers, but this patterns vary among countries (Eurostat, 2019). In this analysis we are going to focus on Spain along a period of time covering the years before and during the financial crisis.

The main objective is to see if there is truly a gender wage gap in Spain, and how it affects women of different ages. Because the gender wage gap is going to be analyzed for three different periods of time, it will allow us to see the main trends and evolution of the wage gap.

The structure of this paper will be as follows: First the data being used will be described. Afterwards the methodology will be detailed. In the following chapters the results obtained will be mentioned and explained thoroughly. The last chapter will be the conclusion were the main findings and their implications will be summarized.

CHAPTER 2. DATA

The wage gap is analyzed by using data from the Encuesta de estructura salarial (salary structure survey) or EES for short that the *Instituto Nacional de Estadística* (INE) publishes every four years (INE, Instituto Nacional de Estadística, 2017). The data goes from 2003 to 2014 and is provided by three quadrennial datasets (years of 2006, 2010, and 2014). The three datasets have 235.272, 216.769, and 209.436 observations respectively. The EES is further consolidated at the European level by using country data provided by all European countries although our analysis focuses only in Spain.

The INE provides microdata for the 3 years. For 2010 and 2014 the INE also provides the code for the data to be imported in R, so the data was easily loaded in this environment. However, the microdata for 2006 was only provided in a .txt format. It was upload to R by using the code provided by Gil (2016).

The EES uses October as the base month in the surveys due to its lack of holidays, season variations, and extra pays. It could be considered as the most "normal" month in a year. There are some aspects to consider for this dataset because not all workers will be taken into account. For instance, professions were salary is mainly obtained from commissions or benefits are not taken into account; and although most economic activities are included in the survey, some such as domestic personnel, agricultural, livestock and fishing activities; partially, the mandatory Public Administration, Defense and Social Security are not included. Also, only workers that are registered in Social Security during the whole month of October in the base year will be included.

The survey provides different types of variables, mainly there are the ones that can affect all workers collectively (and their salaries) such as the market the company operates on, the size of the company, region, etc. There are also variables that give information about the salary such as hours worked, holidays, or duration of the contract. We are going to focus on the variables that are more likely to affect the salary which includes education level, years of seniority, or gender.

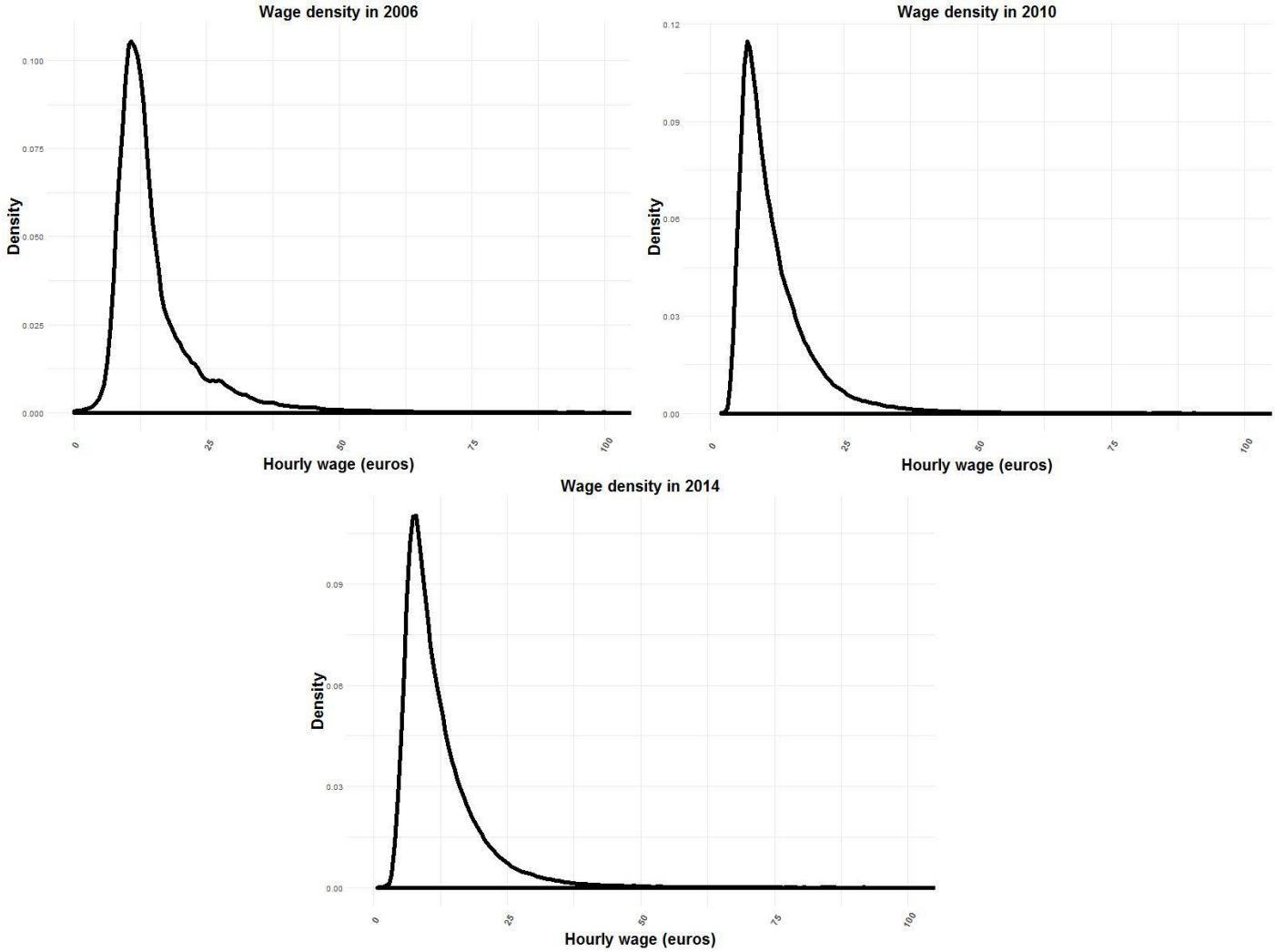
Weighting is important to provide accurate unbiased values. The variable "FACTOTAL" gives weights to each worker in the sample depending on its representation of the whole

population. The weighting factor has been included in the regression. The INE doesn´t provide much information on how it's calculated.

To analyze the wage gap by age and gender the three main variables are hourly salary, gender, and age. Other variables such as education, seniority, and occupation will also be considered, but they will be used as control variables.

Hourly salary is not provided directly in the microdata, so some calculations are required to obtain it, which will be explained in the methodology chapter. The variable is a numerical variable which measures the euros earned for an hour worked. Figure 1 shows how this variable behaves each year. It displays hourly wages from 0 to 100 euros. There is a right skewedness in all the years. It seems to be slightly more skewed after 2006.

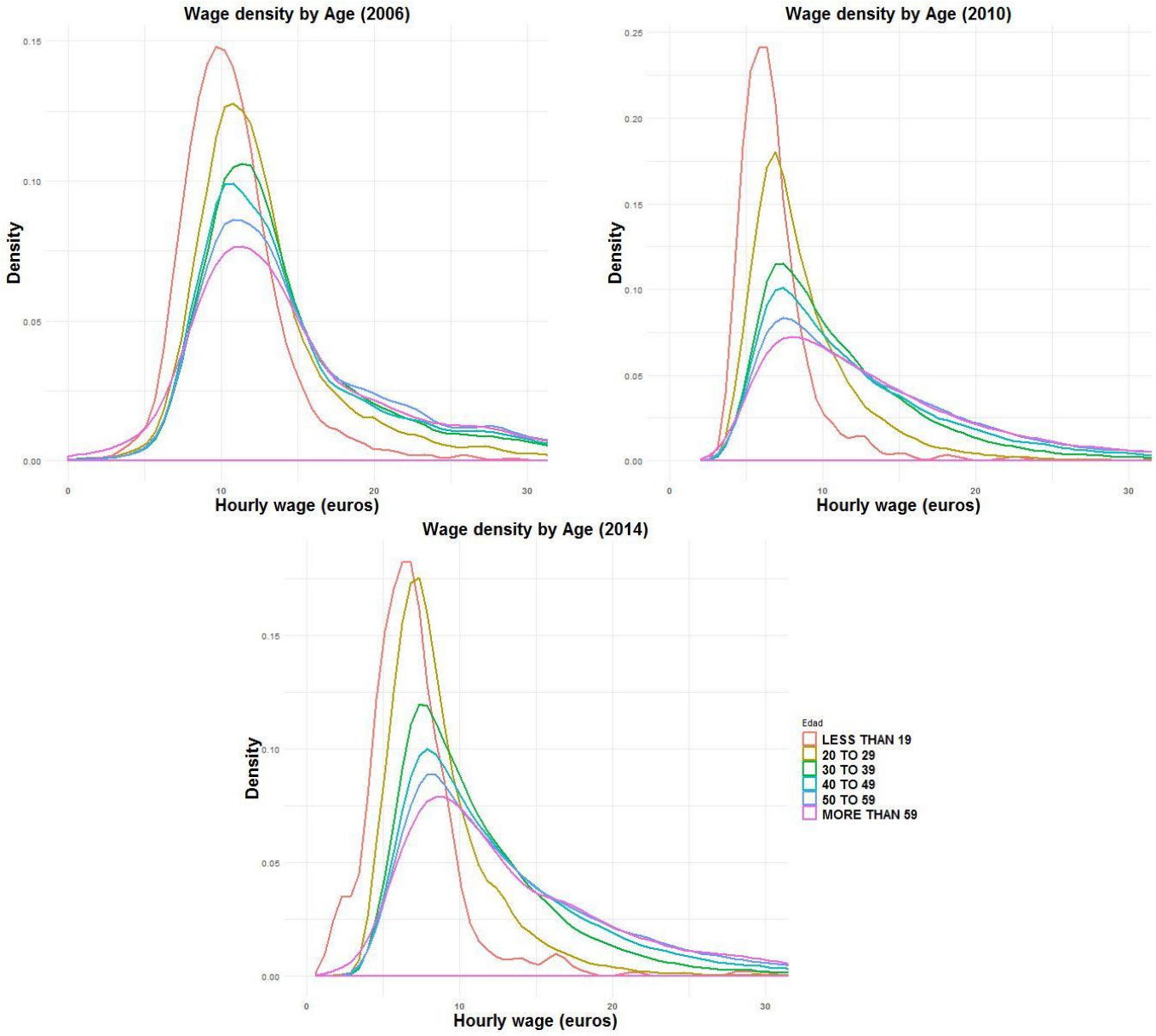
Figure 1. Density distribution of hourly wages in 2006, 2010, and 2014.



Source: Compiled by the author based on (INE, 2019).

The explanatory variables that are used are gender and age. The gender variable is going to be a binary variable that is True when its female and False when its male. For the Age variable, the INE already provided the variable separated into six categories: Less than 19, from 20 to 29, from 30 to 39, from 40 to 49, from 50 to 59, and more than 59 years old. Figure 2 shows the distribution of the hourly salary for each age group. The younger the age group is the more skewed it is, indicating that younger workers have a lower salary than older ones.

Figure 2. Density distribution of hourly wages within age groups in 2006, 2010, and 2014



Source: Compiled by the author based on (INE, 2019).

Three control variables are used to avoid bias, namely Education, Seniority, and Occupation. There are seven levels of education: Less than primary, Primary education, First stage of secondary education, Second stage of secondary education, Advanced vocational training and similar education, University graduates, and University post-graduates and similar, and PhDs. Seniority indicates how long the worker has been with the company he/she is currently working. The dataset provides two variables, one for the years with the company and another for the months; both are numerical variables. Lastly, Occupation is classified according to the National Occupation Classification (CNO by its Spanish acronym). The classification has changed in the different datasets. The one used in 2006 (CNO-94) is different from the one used in 2010 and 2014 (CNO-11). Occupation is being used as a control variable, so we consider unnecessary to homogenize it.

The next step is to analyze and modify the variables to better fit the functional model of the regressions. This is explained thoroughly in the next chapter.

CHAPTER 3. METHODOLOGY

This chapter describes how and why the variables are analyzed. There are three datasets with the same variables from different years, so most of the formulas shown here will be applied equally to all the years.

Beginning with the dependent variable, the hourly salary is not provided directly in the microdata used so it has to be computed. Depending on the year, the way of calculating the salary varied mainly due to the differences of methodology and data acquisition explained before.

For the 2006 dataset the hourly salary is computed with a basic formula that the INE explicitly explains. The hourly wages are estimated as the monthly salary divided by the hours worked (regular and extraordinary) of the reference month:

"la ganancia por hora se ha estimado como la ganancia mensual dividida entre las horas trabajadas (normales y extraordinarias) del mes de referencia." (INE, 2012)

The formula used in 2006 is:

$$SALHORA_{2006} = \frac{SALBASE + EXTRAORM + PHEXTRA}{HEXTRA + JSP1 * 2} \quad (1)$$

For the datasets of 2010 and 2014 the INE provided a document which explains the calculation of the hourly salary. To calculate the hourly salary the salary earned during the month of October must be divided by the hours worked. But factors such as extra hours, how much is paid for those hours or extraordinary payments have to be taken into account (even when using the base month). To begin the calculation, the formulas will take into account the years where $t = 2010, or\ 2014$. We will consider only the days worked (DIASMES), which are calculated through the subtraction of the duration of the work relation in October (DRELABM) minus the days in a "Special situation 2" (DSIESPM2), which will look like:

$$DIASMES_t = DRELABM_t - DSIESPM2_t \quad (2)$$

Afterwards, only the workers that didn't have a "Special situation 1" are included to calculate the components of the monthly salary that are affected by the numbers of days

worked, these are the base salary (SALBASE), the salary allowances (COMSAL), and salary shift work allowances (COMSALTT); the calculation looks like this:

$If\ SIESPM1_t =$ "6". Then do:

$$SALBASE_{2t} = \left(\frac{31}{DIASMES_t}\right) * SALBASE_t \quad (3)$$

$$COMSAL_{2t} = \left(\frac{31}{DIASMES_t}\right) * COMSAL_t \quad (4)$$

$$COMSALTT_{2t} = \left(\frac{31}{DIASMES_t}\right) * COMSALTT_t \quad (5)$$

The other components of the monthly salary not affected by the days worked are extraordinary monthly payment (EXTRAORM) and the payment for extra hours (PHEXTRA). To get the monthly salary (SALMES), the calculation is the sum of all its components:

$$SALMES_t = SALBASE_{2t} + COMSAL_{2t} + COMSALTT_{2t} + EXTRAORM_t + PHEXTRA_t \quad (6)$$

To get the HOURLY salary the monthly salary has to be divided by the hours worked (JMP1), this is obtained by the sum of the hours (JSP1) and minutes (JSP2) stipulated in the weekly workday, plus the extraordinary hours (HEXTRA); to standardize the measure into hours the measurement in minutes is divided by 60 and the sum of the stipulated weekly workday is multiplied by 4.35, which is the average number of weeks a month has. The calculation is:

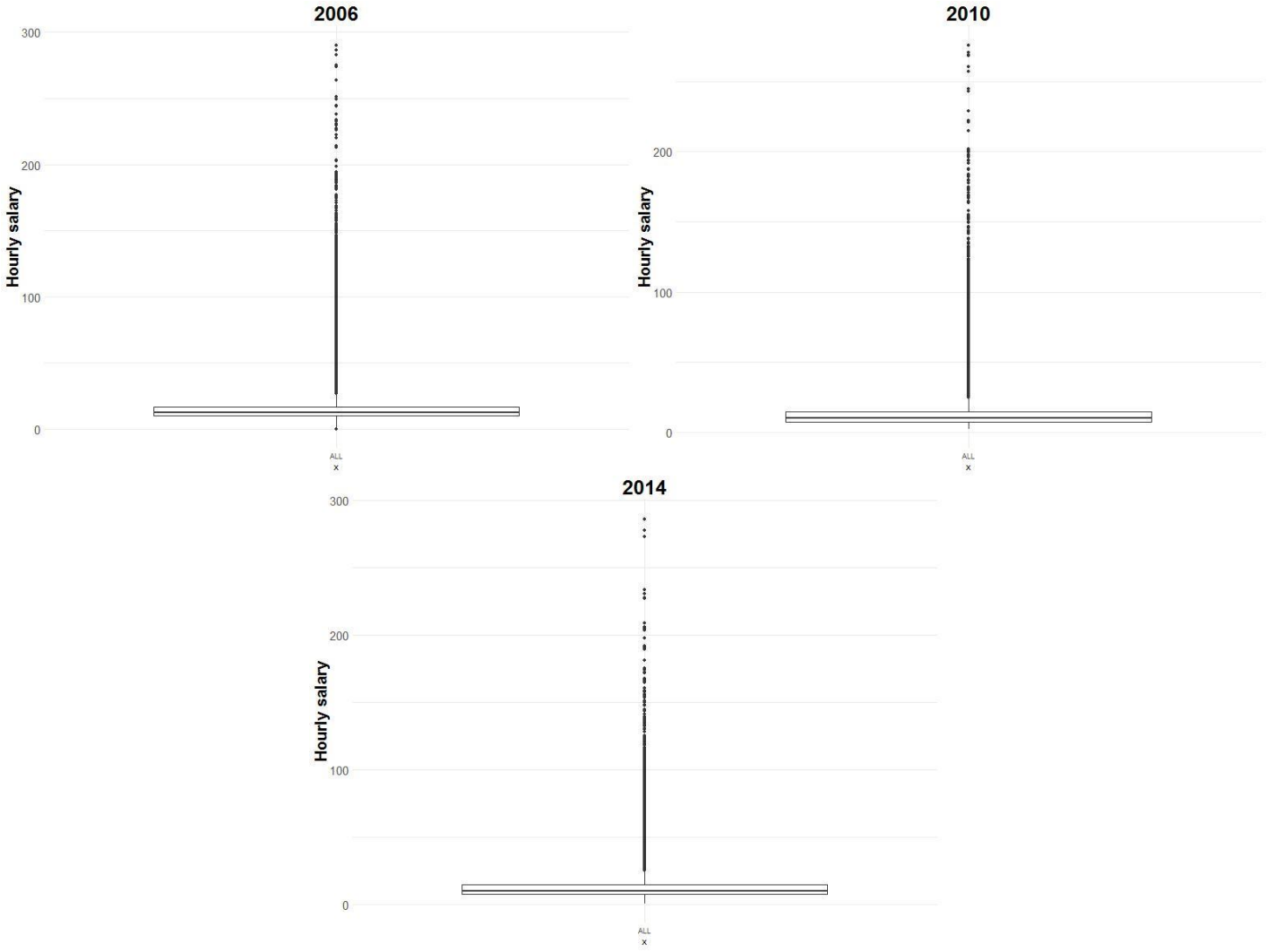
$$JMP1_t = \left(JSP1_t + \frac{JSP2_t}{60}\right) * 4.35 + HEXTRA_t \quad (7)$$

After all these calculations, simply dividing the monthly salary by the hours worked the hourly salary can be obtained:

$$SALHORA_t = \frac{SALMES_t}{JMP1_t} \quad (8)$$

After obtaining the hourly salary, the distribution of hourly salary shows many variables above the 75th percentile. The values of an hourly salary above 300 euros per hour were removed, the results of removing these outliers can be seen in Figure 3.

Figure 3. Quartile distribution of the hourly salary after removing outliers in 2006, 2010 and 2014.

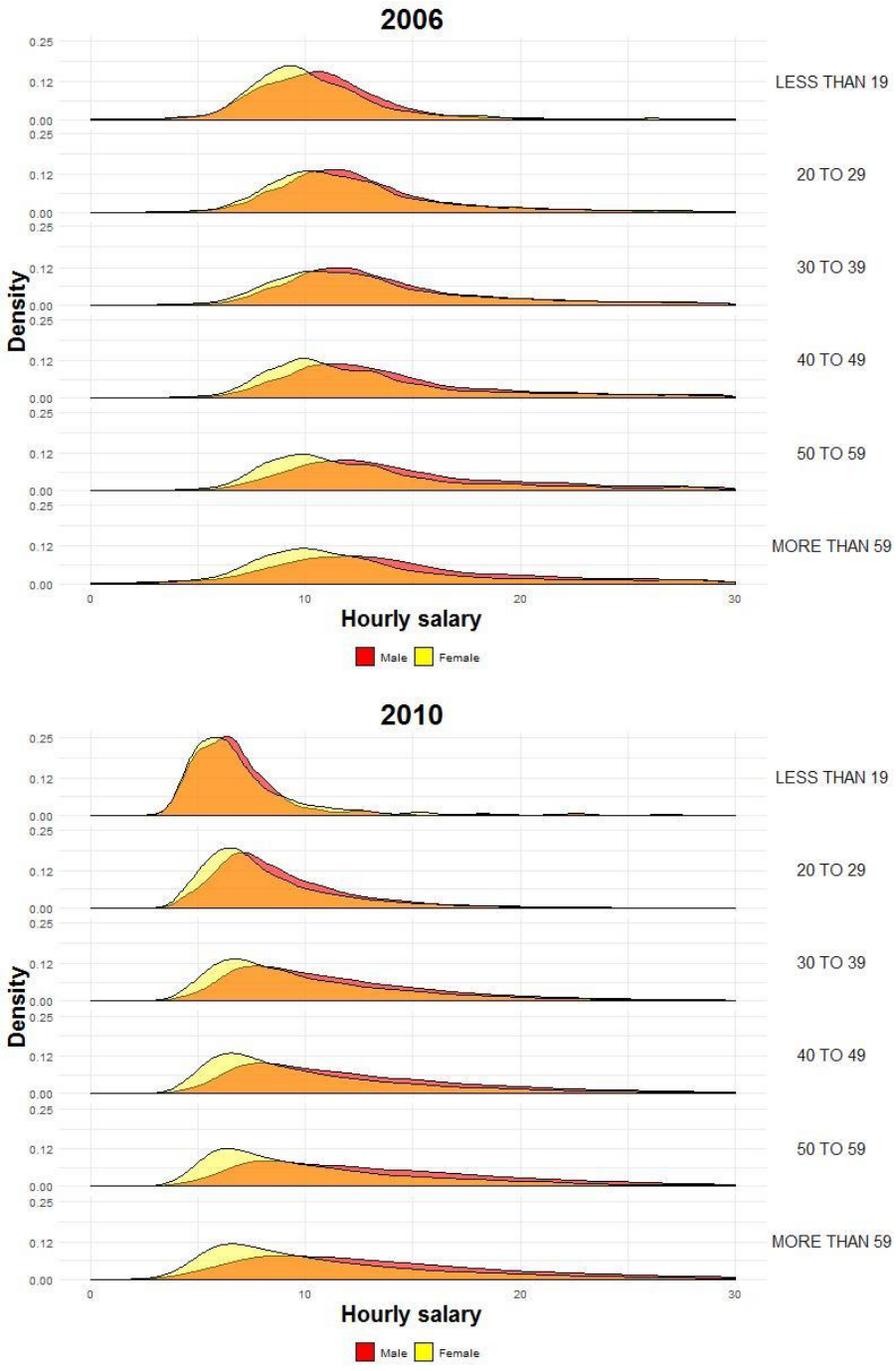


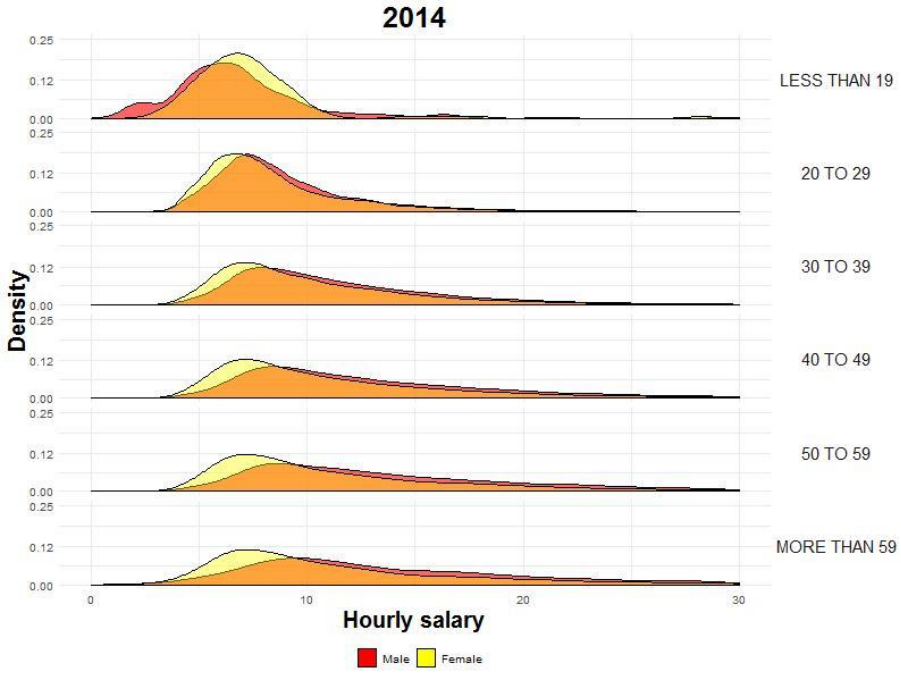
Source: Compiled by the author based on (INE, 2019).

The dependent variable is skewed to the right and truncated at 0, because estimates for hourly wages of 0 or below wouldn't make sense. The variable is transformed into logs as usually described in the literature (Peñas, 2002). After the log transformation, some variables had to be removed because a small number of observations close to 0 transformed into logs will come out as infinite negative. The number of observations removed was 184 in 2001, 0 in 2010, and 1 in 2014.

We use two explanatory variables: Gender and Age. The INE already classifies the Age variable into six bins that have been mentioned before. To describe the variables two types of graphs are provided: density (Figure 4) and quantile (Figure 5) distributions by age and gender.

Figure 4. Comparison of the density distribution of gender among age groups in 2006, 2010 and 2014.





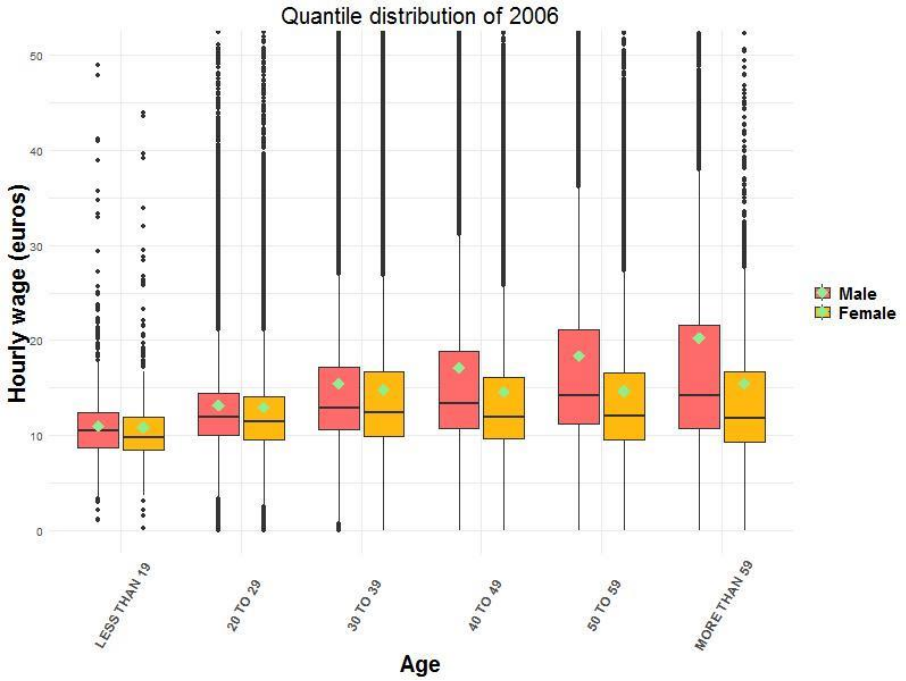
Source: Compiled by the author based on (INE, 2019).

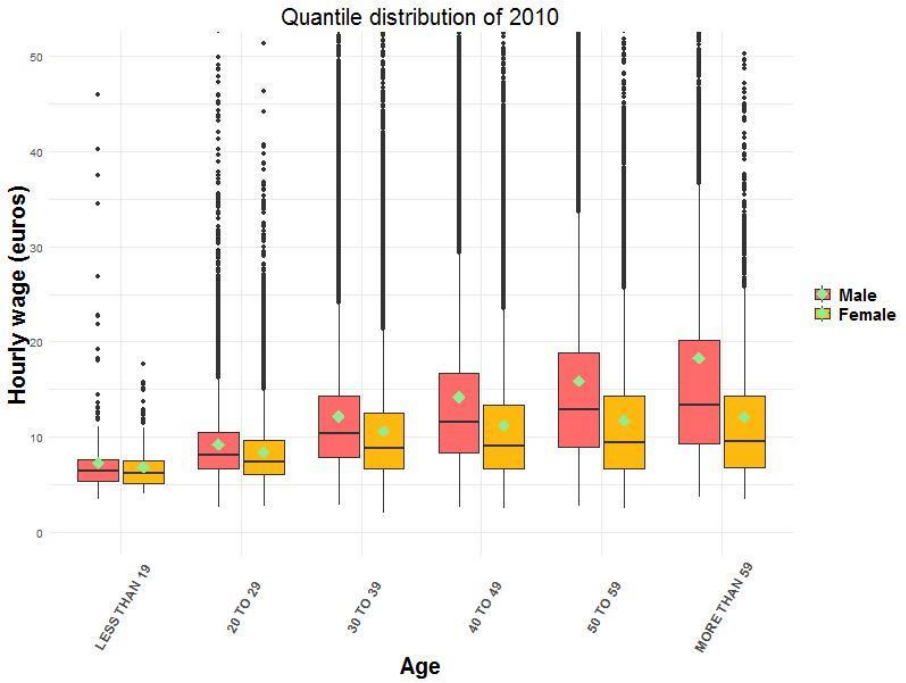
Figure 4 shows the density of males and females in each age group; females are represented by the yellowish shade while males are represented by the red shade. Inside each year, the density is separated into the six age groups to be able to see how each of the groups varies differently throughout the years.

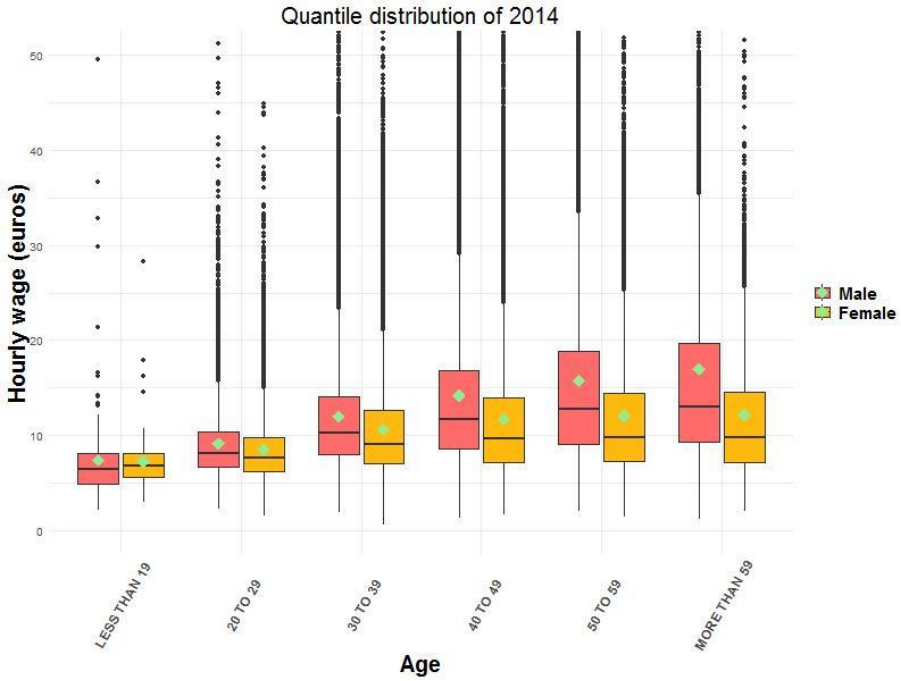
To be able to compare the three years the axes take the same values. The density axis goes from 0 to 0.25 and the hourly salary axis goes from 0 to 30. The latter axis could be up to 300 but shortening it allows a better visualization of the salaries with highest densities. It also depicts better the skewedness of wages.

The graphs show how salary has decreased since 2006 independently of age and gender (Figure 5 will depict this more clearly). Females tend to earn less than males regardless of their age; the "peak" of density for females is to the left of the "peak" of density for males in most cases. There is one exception in 2014 for the youngest age group, less than 19, this group has seen an increase in salary for women compared to men, although both men and women earn less compared to the levels of 2006. It can be seen how the density flattens the older the age group is; also, men tend to earn more the older they get, and the density shifts to the right, while for women the density shifts at a slower pace.

Figure 5. Comparison of the quartile distribution of gender among different age groups in 2006, 2010, and 2014.







Source: Compiled by the author based on (INE, 2019).

In Figure 5 it can be seen how the mean (represented by a green diamond) and the quartiles behave between age groups throughout the three years being analyzed. Females are the yellow color and males the pink one. As in Figure 2 the axes are the same for the three graphs so they can be compared more easily, hourly wage goes from 0 to 50 in this case.

Figure 5 is very similar to Figure 4; both help see the distribution of the salary across ages and gender; but Figure 5 shows more clearly how in general wages decrease after 2006. The means and medians of all ages and both genders were above 10 euros per hour in 2006. In 2010 this only happens for men older than 30 and for women the median never reaches the 10 euros per hour threshold. In 2014 the distribution of wages doesn´t seem to change much, the 25^{th} quartile gets longer and the 75^{th} quartile shrinks; also, the median salary of women tends to increase in all age groups.

The control variables are used to compare males and females of the same age group with similar characteristics (all other things equal or ceteribus paribus); the usefulness of using this control variables will be explained in detail when the data is provided. The three control variables are Education, Occupation, and Seniority. Education is classified into seven bins by the INE. Occupation is classified by the National Occupation Classification

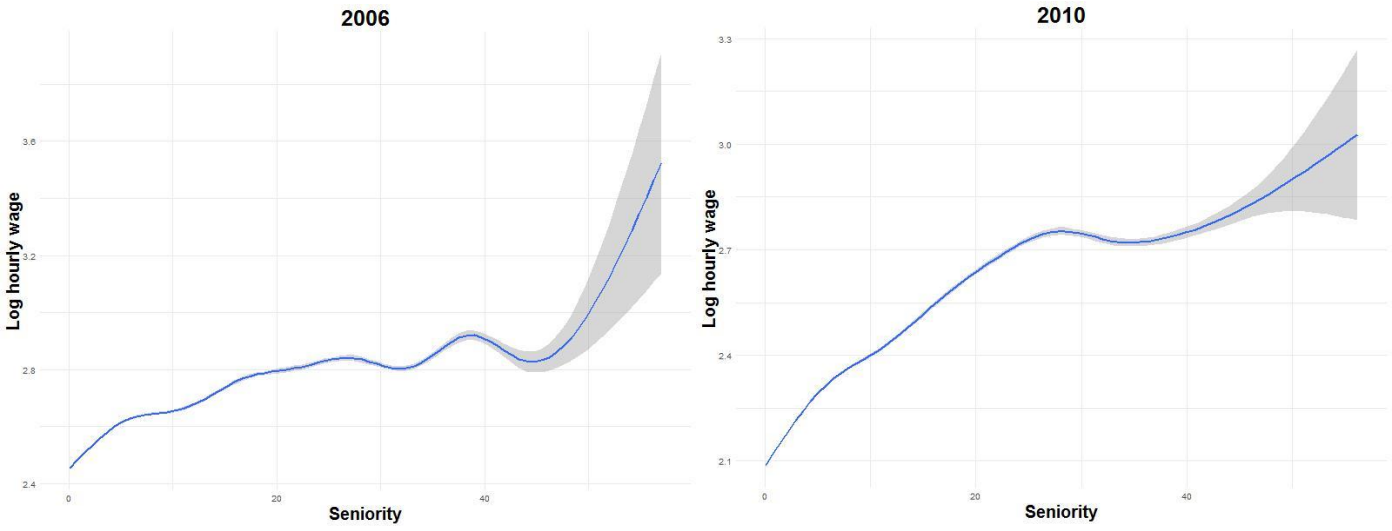
and gives information about the type of job each person has, which allow to compare people in similar positions.

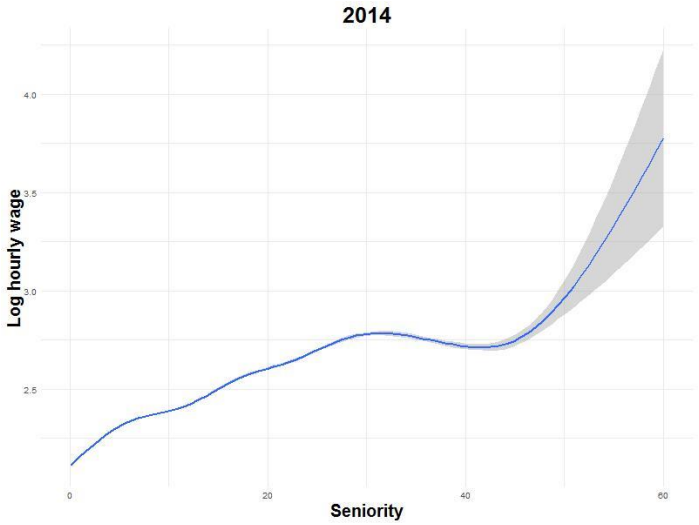
Lastly, as explained in the Data chapter seniority is given in two variables, one with the months and another with the years in the company. Before analyzing the variable, it had to be unified into a standard measurement (years). The formula used was:

$$Seniority = Years_{with\ the\ company} + (\frac{Months_{with\ the\ company}}{12}) \quad (9)$$

Afterwards, as a linear variable, seniority provides the number of years a person has been working in their current workplace. When graphing the smoothed relationship between seniority against hourly salary in Figure 6 there isn't a clear linear relationship, and it varies in each case.

Figure 6. Smoothed relationship of Seniority and log Hourly wage in 2006, 2010, and 2014.





Source: Compiled by the author based on (INE, 2019).

Some authors use seniority squared to assimilate its fluctuations; like Peñas (2002), who also uses the EES for the year of 1998, found some trouble with the seniority variable squared due to some collinearity problems. Other authors separate the variable into groups; like Hernandez Martinez (1995), who separates working experience (which is very similar to seniority) into three groups. The first group is for less than two years, the second group includes those that have worked between two to five years, and the third group includes all the workers with more than five years. In this case, instead of using the variable as linear, seniority is separated into the same three groups as Hernandez.

With all these variables that have been mentioned, there will be two main types of regressions to explain the differences in salary. In all the regressions the explanatory variable will be log transformed, this is because of the large amount of evidence that the relation between the explanatory variables and salary is as a semi elasticity. The interpretation of the coefficients will be in the form of semi elasticities, which will be explained in detail in the Results chapter. The regressions' sub index y will indicate the year the regression belongs to, where $y = 2006, 2010, or\ 2014$.

First there will be the most basic regressions without interactions.

The first one doesn't have controls:

$$Log\ Hourly\ Salary_y = \beta_0 + \beta_1 * Gender_y + \beta_2 * Age_y \quad (10)$$

This regression allows to have a basic view of how gender and age seem to interact with wages by themselves.

Then there are two regressions with different controls:

$$Log\ Hourly\ Salary_y = \beta_0 + \beta_1 * Gender_y + \beta_2 * Age_y + \beta_3 * Education_y + \beta_4 * Seniority_y \quad (11)$$

$$Log\ Hourly\ Salary_y = \beta_0 + \beta_1 * Gender_y + \beta_2 * Age_y + \beta_3 * Education_y + \beta_4 * Seniority_y + \beta_5 * Ocupation_y \quad (12)$$

The only difference between regressions (11) and (12) is that the latter also controls for occupation, a key control variable. This will allow us to compare women and men of the same education level, same years in the company, and similar jobs. These last regressions will give a different comparison of the wage gap than the first regression.

The other type of regressions used will have the interaction of $Gender * Age$. This interaction comes from the hypothesis that the wage gap between men and women depends on their age. Meaning that the wage gap won't be the same for a woman in her 20s than for a woman in her 60s. The regressions will be similar to the ones above, but with the interaction added; which will change the value of the βs.

Without any controls:

$$Log\ Hourly\ Salary_y = \beta_0 + \beta_1 * Gender_y + \beta_2 * Age_y + \beta_3 * (Gender_y * Age_y) \quad (13)$$

This regression is useful to see how the interaction between gender and age affects salary, but there are many unobserved variables that need to be included.

Then there are two regressions with controls and interactions:

$$Log\ Hourly\ Salary_y = \beta_0 + \beta_1 * Gender_y + \beta_2 * Age_y + \beta_3 * \left(Gender_y * Age_y\right) + \beta_4 * Education_y + \beta_5 * Seniority_y \quad (14)$$

$$Log\ Hourly\ Salary_y = \beta_0 + \beta_1 * Gender_y + \beta_2 * Age_y + \beta_3 * (Gender_y * Age_y) + \beta_4 * Education_y + \beta_5 * Seniority_y + \beta_6 * Ocupation_y \quad (15)$$

Regression (15), which includes the interaction and all the control variables being used will give the closer value to the arbitrary wage gap where independently of education, job, and experience, a woman will have a different wage than a man with the same qualifications.

After clarifying how the variables used have been obtained, how they work, and how they are going to be analyzed, the next step is to see what the results will be.

CHAPTER 4. RESULTS

This chapter describes the results of the analysis, namely the main coefficients of the regressions and a comparison between the results of each year. The coefficients can be seen in Tables 1-3. There is a table per year with the results of the six regressions performed for each year. The order of the regressions goes as explained in the Methodology chapter and it's the same for all years; the first three regressions don't have interactions and the last three do.

The sixth regression is the most sophisticated, including all controls and the interaction terms. The base case is a worker male in the age group of 40 to 49 years old. We select that group because it is usually the age with higher wages, making it simpler to compare with the other groups. Lastly, in all regressions the observations are weighted by using the variable provided by the INE, FACTOTAL.

In these regressions without interactions the coefficient of gender is interpreted as the percentage difference of hourly wages between males and females. If negative, females earn less than males and if positive females earn more than males. The age coefficient shows the differences in wages among age groups, it's important to remember that age is separated into six different groups but will only show five because one of the groups is used as the baseline. If the coefficient is negative the age group chosen earns less than the base age group and vice versa. It is important to note that the coefficient of age doesn't distinguish between males and females.

Starting with the simplest regression, all coefficients are statistically significant. The coefficient of female is always negative indicating lower wages for females than for males. The worst year is 2010 with an hourly wage 14.4% lower for females than males, the best was 2006 with just 6.8% less. Because the age variable shows the difference in wages between the age group of choice and the base age group, hourly wages are lower for younger workers than the base age (40 to 49 years) and higher for older workers.

The second regression shows the same coefficients, but controlling for the level of education and the seniority of the workers. The percentage difference in wages between male and females increases when controlling for these variables meaning that both factors differ between male and female workers and are related to the wage. The highest

difference is in 2010, where women earn 18.9% less than men who have the same level of education and have been working in the company the same amount of time. In 2006 the difference is of 11.2% and in 2014 of 18%.

Age acts in a similar way, the younger the worker the lower their salary; but now that education and seniority are controlled for, the size of the coefficient has become smaller. There is one exception in the regressions of 2010 and 2014. It suggests that, when adding control variables, the difference in salary between the base age group and the youngest age group is lower in absolute terms than the difference in salary between the base age group and the age group of 20-29 years. It makes sense if considering that the level of education and the seniority of those groups are likely to be very different.

Moving onto the last regression without interaction, this third regression adds a control variable to the other two used before, which is occupation. In this case, all coefficients are closer to zero, meaning that their effect on hourly wages is lower. The female coefficient still has the highest difference in 2010 with a 15.8% decrease, but compared to the second regression, it has been reduced by three percentage points. It suggests than women tend to work in different jobs — with different wages — compared to men. In the case of the age coefficients, it's the similar as the second regression but with the coefficients closer to zero so the effect of age on wages is lower.

Now, we encounter the regressions with interactions. Their interpretation won't be as straightforward as before. To clarify, the formulas used to calculate the effects of gender and age will be shown.

Effect of being female on hourly salary:

$$\frac{\partial Log\ Hourly\ Salary_y}{\partial Gender_y} = \beta_1 + \beta_3 * Age_y \quad (16)$$

Effect of age on hourly salary:

$$\frac{\partial Log\ Hourly\ Salary_y}{\partial Age_y} = \beta_2 + \beta_3 * Gender_y \quad (17)$$

To interpret the effects of being female on salary will also depend on the age of the worker. We have three types of coefficients, gender, age, and their interaction. The new

set of coefficients from the interaction show the wage gap between males and females of that specific age group; this is very relevant because the wage gap is likely to change with the age of the workers.

For the first regression with interaction no control variables are used. What is particularly relevant is that the coefficients of $Gender_y * Age_y$ where the age is 39 years or less are positive, meaning that the wage gap in those groups is lower, and the older the age group the larger the gap gets. It suggests that the gender gap is decreasing in the younger cohorts.

For regressions (5) and (6) aside of the interaction, there are also control variables. For regression (5) the control variables are Seniority and Education, and for regression (6) the control variables are Seniority, Education, and Occupation. For all three cases the coefficient of gender increases when doing the first control and decreases again when doing the second control.

When including the interactions, for 2006 when increasing the number of variables used as controls the coefficients of age and the interactions tend to decrease towards zero; which means age and gender combined have less relevance on hourly wages. But for 2010, the interaction of female with ages 30-39 is not statistically significant when using all control variables, and only statistically significant for a confidence interval of 95%. In 2014 something similar happens, the coefficients for the interaction of female and ages 50-59 years are only significant at a 95% confidence interval; for the interaction of female and ages 30-39 years when controlling for all variables the coefficient is not significant. Lastly, the coefficient of the interaction for females of ages less than 19 increases when using two control variables and decreases when adding the third, but it is still bigger than when no controls where used.

What we consider the most relevant finding is that the interactions show a positive value for age groups younger than the base case and negative values for age groups older than the base case, meaning that the wage gap is more prominent for older groups. The gender wage gap, although still unbearable, seems to be reducing for younger cohorts. For a better understanding of the coefficients there will be a sub-chapter with a thorough explanation of how to interpret them.

In brief, for the three years two conclusions can be obtained from the coefficients of the interactions: (1) the more control variables used the smaller the wage gap is inside each age group; and (2) the wage gap between males and females increases with age. For this

reason, the most important regression is regression (6), with the interaction and all the controls; and it is the one we are going to focus on from now on to try and extract some meaningful conclusions.

Table 1. Regressions for the year 2006

	Dependent variable:					
	Log Hourly salary 2006					
	No interaction nor controls	No interaction case (1)	No interaction case (2)	Interaction no controls	Interaction case (1)	Interactions case (2)
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	2.633***	2.416***	3.015***	2.648***	2.424***	3.015***
	(0.002)	(0.004)	(0.007)	(0.002)	(0.004)	(0.007)
Female	-0.068***	-0.112***	-0.100***	-0.105***	-0.126***	-0.110***
	(0.002)	(0.002)	(0.002)	(0.004)	(0.003)	(0.004)
Less than 19 years	-0.273***	-0.108***	-0.095***	-0.300***	-0.135***	-0.115***
	(0.008)	(0.008)	(0.007)	(0.011)	(0.010)	(0.009)
20-29 years	-0.124***	-0.066***	-0.052***	-0.164***	-0.088***	-0.066***
	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
30-39 years	-0.019***	-0.032***	-0.025***	-0.049***	-0.044***	-0.034***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
50-59 years	0.043***	0.043***	0.039***	0.068***	0.063***	0.055***
	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)
More than 59 years	0.053***	0.049***	0.044***	0.060***	0.055***	0.048***
	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)	(0.005)
Female*Less than 19 years				0.069***	0.069***	0.051***
				(0.017)	(0.016)	(0.015)
Female*20-29 years				0.091***	0.050***	0.034***
				(0.005)	(0.005)	(0.005)
Female*30-39 years				0.075***	0.029***	0.021***
				(0.005)	(0.005)	(0.004)
Female*50-59 years				-0.074***	-0.057***	-0.048***
				(0.006)	(0.006)	(0.006)
Female*more than 59 years				-0.049***	-0.032***	-0.022**
				(0.012)	(0.011)	(0.010)
Observations	235,178	235,178	235,178	235,178	235,178	235,178
R^2	0.027	0.196	0.247	0.031	0.197	0.247
Adjusted R^2	0.027	0.196	0.247	0.031	0.197	0.247
Residual Std. Error	3.183 (df = 235171)	2.894 (df = 235162)	2.801 (df = 235146)	3.176 (df = 235166)	2.892 (df = 235157)	2.800 (df = 235141)
F Statistic	1,102.074*** (df = 6; 235171)	3,811.872*** (df = 15; 235162)	2,483.431*** (df = 31; 235146)	693.731*** (df = 11; 235166)	2,885.264*** (df = 20; 235157)	2,147.899*** (df = 36; 235141)
Note:	*p<0.1; **p<0.05; ***p<0.01					
	The base age is 40 to 49 years old.					

Case (1) - the regression is controlling for education and seniority.

The base age is 40 to 49 years old.

Case (2) - Case (1) + controlling for occupation

The base age is 40 to 49 years old.

Source: Compiled by the author based on (INE, 2019).

Table 2. Regressions for the year 2010

	Dependent variable:					
	Log Hourly salary 2010					
	No interaction nor controls (1)	No interaction case (1) (2)	No interaction case (2) (3)	Interaction no controls (4)	Interaction case (1) (5)	Interactions case (2) (6)
Constant	2.397***	2.555***	2.923***	2.405***	2.556***	2.921***
	(0.002)	(0.003)	(0.006)	(0.003)	(0.004)	(0.006)
Female	-0.144***	-0.189***	-0.158***	-0.161***	-0.193***	-0.158***
	(0.002)	(0.002)	(0.002)	(0.004)	(0.003)	(0.003)
Less than 19 years	-0.425***	-0.138***	-0.114***	-0.473***	-0.187***	-0.160***
	(0.017)	(0.014)	(0.013)	(0.024)	(0.019)	(0.018)
20-29 years	-0.267***	-0.164***	-0.144***	-0.307***	-0.187***	-0.156***
	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)
30-39 years	-0.075***	-0.084***	-0.073***	-0.098***	-0.089***	-0.075***
	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
50-59 years	0.074***	0.075***	0.069***	0.096***	0.086***	0.076***
	(0.003)	(0.002)	(0.002)	(0.004)	(0.003)	(0.003)
More than 59 years	0.140***	0.131***	0.118***	0.177***	0.161***	0.142***
	(0.005)	(0.004)	(0.004)	(0.006)	(0.005)	(0.005)
Female*Less than 19 years				0.102***	0.103***	0.098***
				(0.034)	(0.028)	(0.026)
Female*20-29 years				0.080***	0.046***	0.023***
				(0.006)	(0.005)	(0.005)
Female*30-39 years				0.049***	0.010**	0.004
				(0.005)	(0.004)	(0.004)
Female*50-59 years				-0.053***	-0.025***	-0.018***
				(0.006)	(0.005)	(0.005)
Female*more than 59 years				-0.105***	-0.082***	-0.065***
				(0.010)	(0.008)	(0.008)
Observations	216,753	216,753	216,753	216,753	216,753	216,753
R^2	0.086	0.384	0.458	0.089	0.385	0.458
Adjusted R^2	0.086	0.384	0.457	0.089	0.385	0.458
Residual Std. Error	3.403 (df = 216746)	2.793 (df = 216737)	2.622 (df = 216722)	3.398 (df = 216741)	2.791 (df = 216732)	2.621 (df = 216717)
F Statistic	3,396.034*** (df = 6; 216746)	9,019.542*** (df = 15; 216737)	6,093.739*** (df = 30; 216722)	1,919.552*** (df = 11; 216741)	6,790.194*** (df = 20; 216732)	5,231.390*** (df = 35; 216717)
Note:	*p<0.1; **p<0.05; ***p<0.01					
	The base age is 40 to 49 years old.					

Case (1) - the regression is controlling for education and seniority.

The base age is 40 to 49 years old.

Case (2) - Case (1) + controlling for occupation

The base age is 40 to 49 years old.

Source: Compiled by the author based on (INE, 2019).

Table 3. Regressions for the year 2014

	Dependent variable:					
	Log Hourly salary 2014					
	No interaction nor controls	No interaction case (1)	No interaction case (2)	Interaction no controls	Interaction case (1)	Interactions case (2)
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	2.418***	2.692***	2.966***	2.428***	2.697***	2.967***
	(0.002)	(0.003)	(0.005)	(0.002)	(0.003)	(0.005)
Female	-0.131***	-0.180***	-0.150***	-0.151***	-0.193***	-0.158***
	(0.002)	(0.002)	(0.002)	(0.004)	(0.003)	(0.003)
Less than 19 years	-0.472***	-0.163***	-0.145***	-0.590***	-0.305***	-0.276***
	(0.021)	(0.017)	(0.016)	(0.031)	(0.025)	(0.024)
20-29 years	-0.273***	-0.168***	-0.149***	-0.330***	-0.216***	-0.184***
	(0.003)	(0.003)	(0.003)	(0.005)	(0.004)	(0.004)
30-39 years	-0.081***	-0.082***	-0.074***	-0.105***	-0.090***	-0.077***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
50-59 years	0.057***	0.074***	0.067***	0.072***	0.079***	0.071***
	(0.003)	(0.002)	(0.002)	(0.004)	(0.003)	(0.003)
More than 59 years	0.094***	0.106***	0.094***	0.124***	0.117***	0.102***
	(0.005)	(0.004)	(0.004)	(0.006)	(0.005)	(0.005)
Female*Less than 19 years				0.222***	0.267***	0.245***
				(0.042)	(0.035)	(0.033)
Female*20-29 years				0.115***	0.096***	0.070***
				(0.007)	(0.005)	(0.005)
Female*30-39 years				0.051***	0.017***	0.006
				(0.005)	(0.004)	(0.004)
Female*50-59 years				-0.033***	-0.012**	-0.011**
				(0.006)	(0.005)	(0.004)
Female*more than 59 years				-0.074***	-0.029***	-0.023***
				(0.009)	(0.008)	(0.007)
Observations	209,421	209,421	209,421	209,421	209,421	209,421
R^2	0.072	0.375	0.441	0.075	0.376	0.442
Adjusted R^2	0.072	0.375	0.441	0.075	0.376	0.442
Residual Std. Error	3.243 (df = 209414)	2.662 (df = 209406)	2.516 (df = 209390)	3.237 (df = 209409)	2.659 (df = 209401)	2.514 (df = 209385)
F Statistic	2,707.095*** (df = 6; 209414)	8,961.373*** (df = 14; 209406)	5,516.414*** (df = 30; 209390)	1,543.633*** (df = 11; 209409)	6,643.200*** (df = 19; 209401)	4,744.183*** (df = 35; 209385)

Note: *p<0.1; **p<0.05; ***p<0.01
The base age is 40 to 49 years old.

Case (1) - the regression is controlling for education and seniority.
The base age is 40 to 49 years old.

Case (2) - Case (1) + controlling for ocupation
The base age is 40 to 49 years old.

Source: Compiled by the author based on (INE, 2019).

INTERPRETATION OF THE COEFFICIENTS

To understand what the coefficients represent it's relevant to mention again that: (i) the dependent variable is in logarithms, so the coefficients have to be transformed to percentages, and (ii) the base case is a 40-49 year old male. To simplify the next explanations, examples will be given from regression (6) Table 1.

The first interpretation is from the coefficient of female. It shows the wage difference between a 40-49 year old female and a 40-49 year old male. For example, when the coefficient is -0.11 it means that in the age group of 40-49 years women's wages are 11% less than men's. The second set of coefficients are those of age. They won't represent the

gender wage gap, but the wage gap of men in different age groups. The coefficients show the wage gap between the age group considered and the 40-49 year old group. For example, if the coefficient for the group of 20-29 years of age is -0.066 then men of 20-29 years of age earn 6.6% less than men of 40-49 years of age.

Lastly, the coefficients of the interactions show the differences in wage gap of a woman of the age group chosen compared to a 40-49 year old woman. For example, if the coefficient is 0.034 for women of the age group 20-29, then that age group will have a wage gap 3.4 percentage points lower than the wage gap in the age group of 40-49.

To see the arbitrary wage gap inside each age group we will need to use equation (16). Still using the same example as before, a woman of 20-29 years of age has a coefficient of -0.11 for being female, and a coefficient of 0.034 for being in the 20-29 year group and being female. The sum of both coefficients will be -0.076. That result shows that the wage gap between males and females of the age group 20-29 will be 7.6%.

To be able to see clearly the wage gap for each age group Table 4 shows the results of the calculations as the one done above for regression (6). A positive number means that there is a wage gap for the specific age group where women are earning less than men, and a negative number is also a wage gap where men are earning less than women. From the table it is easy to see how the wage gap increases with age, and how since 2006 it seems to have increased.

Table 4. Percentage of the wage gap in each age group per year

Age	**2006**	**2010**	**2014**
LESS THAN 19	5,8%	6,1%	-8,7%
20 TO 29	7,6%	13,5%	8,7%
30 TO 39	8,8%	15,4%	15,1%
40 TO 49	11,0%	15,8%	15,8%
50 TO 59	15,7%	17,6%	16,8%
MORE THAN 59	13,1%	22,3%	18,0%

Source: Compiled by the author based on (INE, 2019).

LIMITATIONS

We want to point out what we consider the most relevant limitations of the analysis. First, it is possible that not all the factors related at the same time to the wage of a person and to the explanatory variables have been included in the model and therefore the results may be biased. For instance, the sector is not taken into account but it is relevant considering that "male dominated industries tend to have higher wages" (Vagins, 2019). Also, the nature of the company (private or public) and the localization of the firm will influence this wage gap. Although some of these variables could be find in the datasets, we decided not to include them in the models due to the specific characteristics and limitations of this analysis. Some other explanatory variables that are not available in the datasets used are contextual factors; such as the size of the family or the firm's sensibility to the workers' work-life balance and reconciliation policies (Chinchilla, Poelmans, & León, 2005). This factors have a strong relationship with gender roles and might thin out the importance of a woman's professional career but are not easily measurable (Heredia, Ramos, Sarrió, & Candela, 2002).

Another limitation is related to the different characteristics of the datasets used in the analysis. For instance, the calculation of the hourly wage is different for the 2006 dataset compared to the 2010 and 2014 datasets. It challenges comparing the different years.

CHAPTER 5. DISCUSSION

The results suggest that there are different components related to the gender wage gap. The more control variables added the more arbitrary the wage gap is, meaning that it is going to show the wage gap caused only by the gender and the age of the worker, all other things equal. Specifically, when considering the Occupation control, there is an arbitrary discrimination (female workers tend to earn lower wages compared to male workers when working in the same job) and a job discrimination (women tend to worker in lower paying jobs). The former is provided by the coefficients of the models controlling by Occupation while the latter is related to the coefficients without the control. Therefore, the difference in the coefficients between the regressions with and without the control variable occupation is likely to be related to women having jobs with lower salary. A way of interpreting the result is that women struggle to get higher paying jobs. This is a vertical discrimination where the "proportion of women decreases as you move up the pyramid hierarchy"(Heredia et al., 2002) which is understood as the glass ceiling.

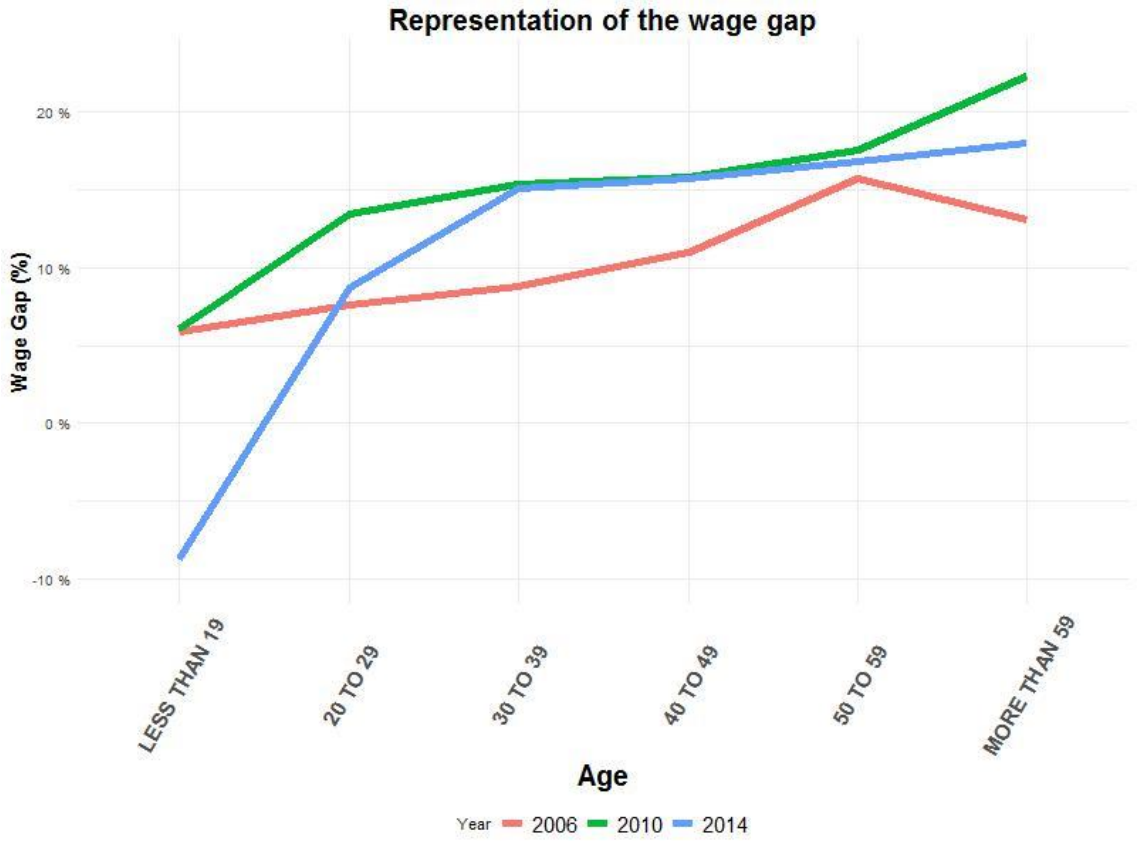
Regarding the change of the gender wage gap by age, after controlling for education, seniority, and occupation, the absolute value of the coefficients increases with the age group. By plotting the coefficients of Table 4 we obtain a visual representation of the wage gap in Figure 7. This figure depicts the value of the gender wage gap, what we call the arbitrary gender gap. A positive value indicates a wage gap were men earn more than women. The higher the percentage the higher the wage gap, and the more men earn compared to women. The positive trend seen when age increases is inherent for the 3 years.

In 2006 the results show a slow increase in the wage gap for younger cohorts, after 30-39 the wage gap increases faster, and for the last cohort (more than 59) the wage gap decreases. This findings are very similar to the ones found by Carnevale et al. (2018). They observed that when looking at the wage gap of college graduates, before reaching the age of 30 the wage gap tends to remain constant. After that, women start to fall behind which causes an increase in the wage gap until its peak at the age of 50-54, after that the gap decreases slightly.

The results in 2014 and 2010 are quite similar. A striking exception appears for the group of less than 19 years of age in 2014; it shows that women's hourly wages were around 8% higher than men's This finding may be related to the low number of observations

(291) of that group compared to several tens of thousands in the other age groups. The difference between 2006 and the other years may be related to an increase in the gender wage gap during the crisis that started in 2008. However, it might also be related to the difference in the estimation of the values that we have already mentioned in the limitations paragraph and that challenges making accurate comparisons.

Figure 7. The gender wage gap of different age groups in 2006, 2010, and 2014.



Source: Compiled by the author based on (INE, 2019).

The results suggest that, although there is a visible wage gap for all workers, younger cohorts tend to enter the workforce with similar salaries, and the difference substantially accentuates for older cohorts (or as they age). In fact, the difference in their hourly salary tends to increase to levels of above 15% and in 2010 it was above 20%. Eurostat (2019) provides some explanations to why the wage gap is more prominent in older age groups at a European level. First, females interrupt their careers once or multiple times which causes them to get behind their male counterparts. And secondly, some equality measures that have been implemented in recent years might not have benefitted older women that already held a job. Another explanation is provided by Carnevale et al. (2018), their main

conclusion is that some of the wage gap is caused by discrimination. This is good and bad news. Although the situation is likely to be improving, particularly for younger women, female workers are still left behind when they advance in their professional career.

FURTHER RESEARCH

It would be interesting to make a deeper analysis of the gender wage gap to find their different components and causes. The first suggestion is to include additional control variables (such as the ones mentioned in the Limitations). Understanding why the wage gap is increasing by age is also another area of further research. Although we have quantified the phenomenon, we are far from knowing the causes. A deep understanding could improve the definition of gender policies in the labor market.

Another interesting analysis that could be done using the Oaxaca decomposition (Hlavac, 2018a), this decomposition is another way of analyzing wage gap. It has been used in many papers like Peñas' (2002), who used the dataset provided in 1995 by the INE.

CHAPTER 6. CONCLUSION

This paper is based on data provided by the INE (2019) Quadrennial Salary Structure Survey to analyze how the gender wage gap varies with age. The dataset includes a big number of observations and relevant explanatory variables that makes analyzing the gender gap feasible. The dependent variable (hourly wages) was transformed into logarithm form to model correctly the effects of the independent variables as suggested by previous research. The two independent variables used were gender and age. Lastly, to allow for a more precise result three control variables were used: education, seniority, and occupation.

The results show how women, on average, face some resistance when advancing their professional careers and might not reach higher wages compared to men. We consider that there are different components related to the gender wage gap. We have detected two of these components in this analysis. One is due to women working on lower paying jobs and the other is what we called the arbitrary gender gap. This latter gap, unfair, means that a woman gets a lower wage than a man when working exactly in the same job. Both components exist and have been assessed in our analysis. Regarding the arbitrary gender gap we have made the analysis of how this gap changes by Age. Our results suggest that, although females earn less than males in each age group, the older the age group the bigger the wage gap. It may be related to different factors, such as older women not benefiting of gender policies, or to the "glass ceiling". Comparing the wage gap among the three years analyzed (2006, 2010, and 2014) 2006 had the lowest gender wage gap, while, during the crisis the wage gap seems to be increasing. However, the differences in the methodology of the different datasets challenges making accurate comparisons.

To conclude, we have seen how the wage gap affects women throughout their whole working lives and how in the past decade it hasn't improved.

BIBLIOGRAPHY

Carnevale, A. P., Smith, N., & Gulish, A. (2018). Women can't win: Despite making educational gains and pursuing high-wage majors, women still earn less than men.

Chinchilla, N., Poelmans, S., & León, C. (2005). Mujeres directivas bajo el techo de cristal. *Directivas En La Empresa: Criterios De Decisión Y Valores Femeninos En La Empresa,*

Eurostat. (2019). Gender pay gap statistics . Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php/Gender_pay_gap_statistics#Gender_pay_gap_much_lower_for_young_employees

Gil Bellosta, C. J. (2016). *MicroDatosEs: Utilities for official spanish microdata* Retrieved from https://CRAN.R-project.org/package=MicroDatosEs

Hegewisch, A., Liepmann, H., Hayes, J., & Hartmann, H. (2010). Separate and not equal? gender segregation in the labor market and the gender wage gap. *IWPR Briefing Paper, 377*

Heredia, E. B., Ramos, A., Sarrió, M., & Candela, C. (2002). Más allá del «techo de cristal» diversidad de género. *Revista Del Ministério De Trabajo E Assuntos Sociales, 40*, 55-67.

Hernandez Martínez, P. J. (1995). *Análisis empírico de la discriminación salarial de la mujer en españa*

Hlavac, M. (2018a). Oaxaca: Blinder-oaxaca decomposition in R. *SSRN Electronic Journal,* doi:10.2139/ssrn.2528391

Hlavac, M. (2018b). *Stargazer: Well-formatted regression and summary statistics tables*. Bratislava, Slovakia: Central European Labour Studies Institute (CELSI). Retrieved from https://CRAN.R-project.org/package=stargazer

INE. (2012). *Encuesta de estructura salarial (EES) metodología*

INE. (2019). INEbase/ mercado laboral/ salarios y costes laborales / encuestas de estructura salarial /resultados /microdatos. Retrieved from https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177025&menu=resultados&secc=1254736195110&idp=1254735976596

INE, Instituto Nacional de Estadística. (2017). *Encuesta de estructura salarial (EES) metodología*. Madrid:

Kessler, G. (2014). President obama's persistent '77-cent' claim on the wage gap gets a new pinocchio rating. Retrieved from https://www.washingtonpost.com/news/fact-checker/wp/2014/04/09/president-obamas-persistent-77-cent-claim-on-the-wage-gap-gets-a-new-pinocchio-rating/

Peñas, I. L. (2002). La discriminación salarial por razones de género: Un análisis empírico del sector privado en españa. *Reis,* (98), 171-196. doi:10.2307/40184443

R CoreTeam. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rosin, H. (2013). You know that "Women make 77 cents to every man's dollar" line? it's not true. Retrieved from https://slate.com/human-interest/2013/08/gender-pay-gap-the-familiar-line-that-women-make-77-cents-to-every-mans-dollar-simply-isnt-accurate.html

RStudio Team. (2016). RStudio: Integrated development environment for R. Retrieved from http://www.rstudio.com/

Vagins, D. J. (2019,). The simple truth about the gender pay gap. Retrieved from https://www.aauw.org/research/the-simple-truth-about-the-gender-pay-gap/

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis* Springer-Verlag New York. Retrieved from http://ggplot2.org

Zeileis, A., & Grothendieck, G. (2005). Zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software, 14*(6), 1-27. doi:10.18637/jss.v014.i06

Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News, 2*(3), 7-10. Retrieved from https://CRAN.R-project.org/doc/Rnews/